Further development and applications of Geochip 3.0 for microbial community analysis

Zhili He[1,5], Ye Deng[1,5], Joy D. Van Nostrand[1,5], Liyou Wu[1,5], Christopher Hemme[1,5], Terry J. Gentry[2], Adam P. Arkin[3,5], Terry C. Hazen[4,5], and Jizhong Zhou[1,5]

[1]The University of Oklahoma, Norman, OK, [2]Texas A&M University, College Station, TX, [3]Physical BioScience Division, Lawrence Berkeley National Laboratory, Berkeley, CA, [4]Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, [5]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

ESPP2 Environmental Stress Pathway Project

DOE GENOMICS:GTL ACCELERATING DISCOVERY FOR ENERGY AND ENVIRONMENT

VIMSS Virtual Institute for Microbial Stress and Survival

http://vimss.lbl.gov/

ABSTRACT

Microarray technology provides the opportunity to identify thousands of microbial genes or populations simultaneously. Recently, a comprehensive functional gene array, called GeoChip 2.0, has been developed, evaluated and applied for characterizing microbial communities in natural systems. GeoChip 2.0 contains 24,243 oligonucleotide (50mer) probes and covers > 10,000 genes in >150 functional groups involved in nitrogen, carbon, sulfur and phosphorus cycling, metal reduction and resistance, and organic contaminant degradation. It is a powerful generic tool, and can be used for: (i) profiling various environmental samples, such as soil, groundwater, sediments, oil fields, deep sea, animal guts, and etc; (ii) studying biogeochemical processes and functional activities of microbial communities important to human health, agriculture, energy, global climate change, ecosystem management, and environmental cleanup and restoration; (iii) exploring direct linkages of microbial genes/populations to ecosystem processes and functions; and (iv) detecting functional genes and/or organisms in a particular environment. Here, we present an application example on the dynamics and stability of microbial genes and associated communities during a bioremediation period at the Oak Ridge Field Research Center (FRC). Due to exponential increases in the number of genes and the number of sequences for each gene, a new generation of such an array (GeoChip 3.0) is in development. GeoChip 3.0 is expected to have more features: (i) It is more comprehensive, covering >46,000 gene sequences of 292 gene families; (ii) it includes the phylogenetic marker, *gyrB*; (iii) It is more automatic for sequence retrieval and selection, probe design and verification, array construction and data analysis, information storage, and automatic update, which greatly facilitate the management of such a complicated array, especially for future updates.

Current version: GeoChip 2.0

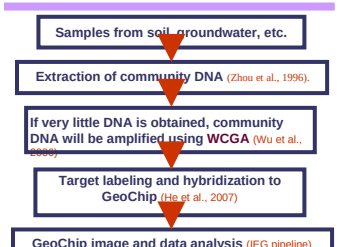
Table 1 List of major functional markers on the GeoChip 2.0

Gene category	Example of key enzyme (gene)	Total probes
Nitrogen cycling		**5310**
Nitrogen fixation	Nitrogenase (*nifH*)	1225
Denitrification	Nitrate reductase (*narG, napA, nasA*), nitrite reductase (*nirS, nirK*), nitric oxide reductase (*norB*), nitrous oxide reductase (*nosZ*)	2306
Nitrification	Ammonium monooxygenase (*amoA*), hydroxylamine oxidoreductase (*hao*)	347
Nitrogen mineralization	Urease (*ureC*), glutamate dehydrogenase (*gdh*)	1432
Carbon cycling		**4599**
Carbon fixation	Rubisco (*cbbL, rbcL*), *Acl* (*aclB*), CODH, FTHFS	1018
Cellulose degradation	Cellulase, endoglucanase	1285
Lignin degradation	Laccase, mannanase	513
Chitin degradation	Endochitinase (*chiA*), exochitinase	744
Methane production	Methyl coenzyme M reductase (*mcrA*)	437
Methane oxidation	Methane monooxygenase (*pmoA*)	336
Others	Lignin peroxidase (*lip*), pectinase, cellobiase	266
Sulfur reduction	Sulfite reductase (*dsrA/B*), APS (*aprA*)	**1615**
Phosphorus utilization	Exopolyphosphatase (*ppx*), phytase	**145**
Metal reduction and resistance		**4546**
Arsenic resistance	Arsenate reductase (*arsC, arsB, arsC*)	877
Cadmium resistance	Cadmium transporter (*cadA, cadB, cadC*)	282
Chromium resistance	Chromium/chromate transporter (*chrA*)	319
Mercury resistance/reduction	Mercuric ion reductase/transporter (*mer, merA, merB*)	548
Nickel resistance	Nickel transporter (*nccA*), permease (*norB*)	140
Zinc resistance	Zinc resistance protein (*znuA*)	128
Other metal resistance/reduction	cobalt resistance proteins, selenium reductase, etc.	2252
Contaminant degradation		**8028**
Benzene, toluene, ethylbenzoic and xylene (BTEX) & related aromatics	Benzene 1,2-dioxygenase (*bnz*), ethylbenzene dehydrogenase (*ebd*), benzylsuccinate synthase (*bss*), xylene monooxygenase (*xyl*), benzoyl-CoA reductase (*bcrA*), and catechol 1,2-dioxygenase (*cat, tfd*).	4176
Chlorinated aromatics	Chlorophenol reductive dehalogenase (*cpr*)	90
Nitroaromatics	Nitrobenzene nitroreductase (*nbr*), 4-nitrobenzaldehyde dehydrogenase (*nmr*), p-nitrobenzoate reductase (*pnb*)	152
Polycyclic aromatic hydrocarbons (PAHs)	Naphthalene dioxygenase (*nah*), PAH ring-hydroxylating dioxygenase (*pdo*)	741
Polychlorinated biphenyls (PCBs)	Biphenyl dioxygenase (*bph*)	388
Chlorinated solvents (e.g. PCE)	PCE/TCE reductive dehalogenase (*rdh, pceA, tceA*)	232
Other organic compounds/by-products	Alkane hydroxylase (*alkB*), homogentisate 1,2-dioxygenase (*hmg*), vanillate O-demethylase oxygenase (*van*), etc.	2249
Total		*24243*

GeoChip design strategies

- Using MSA to identify conserved regions for each functional gene.
- Using experimentally established probe design criteria and the novel software tool *CommOligo*.
- Designing gene-specific and group-specific probes.
- Multiple probes for each sequence or each group of sequences.

Fig. 1 GeoChip analysis of microbial communities



CONCLUSIONS

1. **GeoChip has been constructed with more than 24,000 oligos covering more than 10,000 gene sequences. To our knowledge, this is the most comprehensive functional gene array currently available for environmental studies.**
2. **GeoChip has been evaluated, and demonstrates that it can be used as a powerful tool for a rapid, high-through-put and cost-effective analysis of microbial communities.**
3. **Microbial activities and associated communities were successfully monitored for *in situ* bioremediation at the Oak Ridge FRC site.**
4. **A new generation of GeoChip (version 3.0) with more features is in development, which is expected to provide a more comprehensive picture for a given microbial community.**

ACKNOWLEDGEMENT

ESPP2 (MDCASE) is part of the Virtual Institute for Microbial Stress and Survival (VIMSS) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program:GTL through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

Further Development of GeoChip 3.0

New features for GeoChip 3.0

1. GeoChip 3.0 is more comprehensive, and it contains >24,500 probes and covers about 47,000 (~10,000 for GeoChip 2.0) gene sequences of 292 gene families (~150 gene family on GeoChip2.0) (**Table 2**). Thus, GeoChip 3.0 will be more representative.
2. The homology of automatically retrieved sequences by key words is verified by HUMMER using seed sequences so that unrelated sequences ca be removed.
3. A software package (including databases) has been developed for sequence retrieval, probe and array design, probe verification, array construction, array data analysis, information storage, and automatic update, which greatly facilitate the management of such a complicated array, especially for future updates (**Fig. 2**).
4. GeoChip has implemented a universal standard, which can compare different samples, and normalize data.
5. GeoChip 3.0 implements a genomic control/standard, which can quantitatively analyze functional gene data.

- Automatic update greatly facilitates the management of such a complicated functional gene array.

Fig. 2 Work flow for GeoChip 3.0 design, construction and data analysis

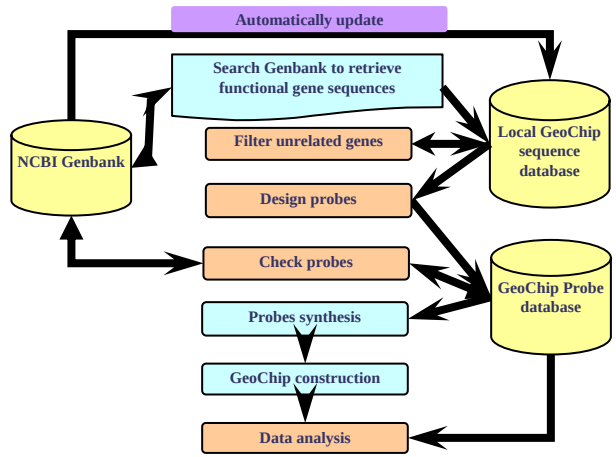


Table 2. The summary of GeoChip 3.0 probe and sequence information

Gene category	No. of gene categories	No. of sequences for probe design	Total no. of probes designed	Total no. of CDS covered
Carbon degradation	31	9839	2720	4737
Carbon fixation	5	3378	898	1806
Methane reduction and oxidation	3	4182	254	434
Metal resistance and reduction	41	16825	4917	10458
Nitrogen cycling	13	27162	3561	6892
Organic remediation	190	31236	8815	16948
Phosphorus utilization	3	1441	599	1212
Sulfur cycling	3	4296	1328	1773
Energy process	2	901	413	449
Others (e.g. *gyrB*)	1	7957	1164	2251
Total	292	107217	24669	46960

Monitoring microbial activities during in situ bioremediation of uranium at the FRC site in Oak Ridge using GeoChip 2.0

Fig. 3 **The microbial community dynamics from the monitoring well (102-2) during the bioremediation period from day 166 (3/1/2004) to day 719 (8/31/2005).**

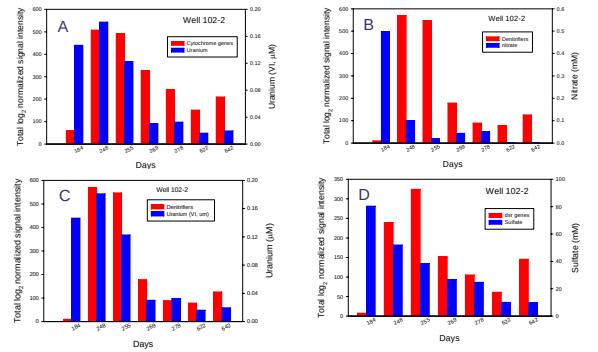
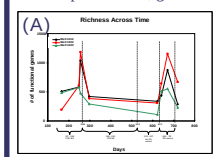
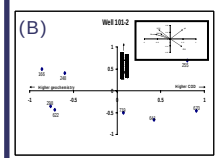
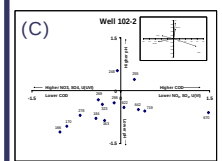
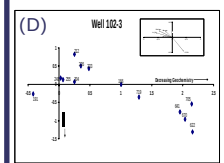


Fig. 4 Statistical analysis of richness across operational time and correlations (A) between GeoChip 2.0 results, geochemistry data, time and microbial communities (B, C, D).









(B) CCA analysis for Well 101-2

Geochemistry: **(i)** Nitrate, COD, and sulfate were most highly correlated to axis 1, and **(ii)** pH, U(IV) were positively correlated to axis 2 (pH more strongly than U(IV)).

Time/communities: **(i)** Day 166, 248, 298, and 622 were all associated with increasing nitrate and sulfate, **(ii)** Day 166 and 248 were also associated with increasing U(VI), **(iii)** Day 255 was associated with pH and COD, and **(iv)** Days 641, 670, and 719 were strongly associated with COD.

(C) CCA analysis for Well 102-2

Geochemistry: **(i)** Nitrate and U(VI) were both strongly correlated to the first axis (negatively). However, they diverged from each other along the gradient; **(ii)** COD again was highly positively correlated to the first axis; and **(iii)** pH and sulfate were both highly correlated to second axis.

Time/communities: **(i)** The early time periods (with the exception of 248 and 255) were distributed along a gradient of nitrate and U(VI), **(ii)** Days 248 and 255 were more correlated to sulfate and pH (primarily pH), **(iii)** The later time periods were distributed along a gradient of COD; and **(iv)** there is a temporal gradient that is likely being strongly influenced by the decreasing nitrate and U(VI) and increasing COD.

(D) Analysis for Well 102-3

Geochemistry: **(i)** All but the COD were negatively associated with the first axis, and **(ii)** COD was highly negatively correlated to the second axis.

Time/communities: **(i)** The gradient of communities is moving away from the geochemistry and toward the COD, **(ii)** Day 191 is highly correlated to pH and nitrate, **(iii)** Day 166, a bit of an outlier, is highly correlated to axis 1 and is being pulled by very high nitrate, but low U(VI) and pH, and **(iv)** all the other communities are along a changing gradient beginning with high geochemistry and low COD, to the later time points where geochemistry is much lower and COD much higher.